THE SINGLE SERVER QUEUE AND THE STORAGE MODEL: LARGE DEVIATIONS AND FIXED POINTS

MOEZ DRAIEF

We consider the coupling of a single server queue and a storage model defined as a queue/store model. We establish that if the input variables, arrivals at the queue and store, satisfy large deviations principles and are linked through an *exponential tilting*, then the output variables (departures from each system) satisfy large deviations principles with the same rate function.

1. Introduction

A celebrated theorem of Burke [1] asserts that when a Poisson process $\{a_n,\ n \in \mathbb{Z}\}$ with mean interarrival time $1/\lambda$ is input to a single server queue whose service times $\{s_n,\ n \in \mathbb{Z}\}$ are i.i.d. exponentials (./M/1/∞ following the Kendall nomenclature) with mean $1/\mu < 1/\lambda$, the equilibrium departure process $\{d_n,\ n \in \mathbb{Z}\}$ from the queue is also a mean $1/\lambda$ Poisson process. In this sense the mean $1/\lambda$ Poisson process is a *fixed point* for the ./M/1/∞ operator. In [6, 7], we prove a generalization of Burke's theorem to the couple of departures and the sequence of times spent at the very back of the queue $\{r_n,\ n \in \mathbb{Z}\}$. The sequence $\{r_n,\ n \in \mathbb{Z}\}$ can also be seen as the sequence of departures from a storage model, where s_n is the amount of P supplied at slot $n+1$, and a_n is the amount of P asked for at the same slot. More precisely, we show that under the conditions of Burke's theorem the couple of *output variables* from the Queue/Store model $\{(d_n, r_n),\ n \in \mathbb{Z}\}$ has the same law as the couple of *input variables* $\{(a_n, s_n),\ n \in \mathbb{Z}\}$.

In this paper, we consider the fixed point question at large deviations scaling. A similar result has been presented by Ganesh et al. [10] for discrete-time queues. Assuming that the input variables satisfy a large deviations principle, we show that if their laws are linked through an exponential tilting, then the large deviations principle is preserved by the Queue/Store system (i.e., the output variables satisfy large deviations principles with the same rate functions as the input variables).

This paper is organized as follows. Section 2 is devoted to giving some background on large deviations that will be useful for our analysis. The Queue/Store model is presented in Section 3. Section 4 focuses on the workload process. We define the concept of effective bandwidth in Section 5. Finally, in Section 6, we derive a large deviations principle for the output variables of the Queue/Store model (Section 6.1) and we give the condition on the rate functions of the input variables (Section 6.2) which are preserved by the Queue/Store operator.

2. Large deviations

Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $x = \{x_n, n \in \mathbb{N}\}$ be a sequence of i.i.d. random variables. We define the sequence of partial sums $X = \{X_n, n \in \mathbb{N}\}$ by

$$X_n = \sum_{i=1}^{n} x_i. \tag{2.1}$$

If $\mathbb{E}|x_0| < \infty$, then X satisfies *the strong law of large numbers*, that is,

$$\lim_{n \to \infty} \frac{X_n}{n} = \mathbb{E}x_0, \quad \text{a.s.} \tag{2.2}$$

We also focus on the fluctuation of a random variable around its mean. Suppose $\mathbb{E}x_0^2 < \infty$ then the *central limit theorem* gives the fluctuations of the scale of $O(1/\sqrt{n})$, of X_n/n around $\mathbb{E}x_0$. More precisely the sequence $\sqrt{n}(X_n/n - \mathbb{E}x_0)$ converges in law a Gaussian random variable with mean zero and variance the one of x_0. Large deviations theory deals with the fluctuations of the scale of $O(1)$.

Definition 2.1. Let $\mathscr{X}$ be a real Hausdorff space. A function $I : \mathscr{X} \to \overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ is a *rate function* if I is lower semicontinuous, that is, the sets $\{x : I(x) \le \alpha\}$ are closed, for all $\alpha \in \mathbb{R}$. In addition, if these sets are compact, then I is a *good* rate function.

Definition 2.2. A sequence $\{x_n, \ n \in \mathbb{N}\}$ in $\mathscr{X}$ satisfies a *large deviations principle* with the rate function $I : \mathscr{X} \to \overline{\mathbb{R}}_+$, if for each Borel subset $\mathscr{B}$ of $\mathscr{X}$,

$$- \inf_{x \in \mathscr{B}^{\circ}} I(x) \le \liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(x_n \in \mathscr{B}) \le \limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(x_n \in \mathscr{B}) \le - \inf_{x \in \overline{\mathscr{B}}} I(x), \tag{2.3}$$

where $\mathscr{B}^{\circ}$ denotes the interior of $\mathscr{B}$ and $\overline{\mathscr{B}}$ its closure.

Throughout the rest of this section, we are interested in real-valued random variables.

Definition 2.3. The *cumulant generating function* of a real-valued random variable x is given by

$$\Lambda_X(\theta) = \log \mathbb{E}e^{\theta x} \quad \forall \theta \in \mathbb{R}, \tag{2.4}$$

which can be infinite.

Using Hölder's inequality and Fatou's lemma, we check that Λ_X is convex, lower semi-continuous and for all θ in the effective domain of Λ_X (i.e., $\{\theta \mid \Lambda_X(\theta) < \infty\}$), we have

$$\Lambda_X'(\theta) = \frac{\mathbb{E}(X_0 e^{\theta X_0})}{e^{\Lambda_X(\theta)}}. \tag{2.5}$$

We define the *Legendre transform* of Λ_X as

$$I_X(x) = \sup_{\theta \in \mathbb{R}} \{\theta x - \Lambda_X(\theta)\}. \tag{2.6}$$

The function I_X is positive, convex, and lower semicontinuous. If Λ_X is finite at the origin, then

$$\Lambda_X'(0) = \mathbb{E}X_0, \qquad I_X(\mathbb{E}X_0) = 0. \tag{2.7}$$

Moreover, we easily check that

$$\begin{aligned}
I_X(x) &= \sup_{\theta \geq 0} \{\theta x - \Lambda_X(\theta)\} \quad \text{for } x \geq \mathbb{E}x_0, \\
I_X(x) &= \sup_{\theta \leq 0} \{\theta x - \Lambda_X(\theta)\} \quad \text{for } x \leq \mathbb{E}x_0.
\end{aligned} \tag{2.8}$$

We recall the statement of *Cramér's theorem* for real-valued random variables.

THEOREM 2.4. *Let $\{x_n, \; n \in \mathbb{N}\}$ be an i.i.d. sequence of real-valued random variables and let Λ_X be its cumulant generating function which is finite at the origin. Then the sequence of random variables $\{X_n/n, \; n \in \mathbb{N}^*\}$, where $X_n = \sum_{i=1}^n x_i$, satisfies a large deviations principle with rate function I_X defined in (2.6): for each closet subset F of $\mathbb{R}$,*

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \in F) \leq -\inf_{x \in F} I_X(x), \tag{2.9}$$

and for each open subset O of $\mathbb{R}$,

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \in O) \geq -\inf_{x \in O} I_X(x). \tag{2.10}$$

The inequalities (2.9) and (2.10) are known as the upper and lower bounds of the large deviations principle. For the different extensions of Cramér theorem we refer to [5]. We can have a large deviations principle for a couple of independent random variables.

THEOREM 2.5. *Let $\mathscr{X}$ and $\mathscr{Y}$ be two real Hausdorff spaces, and $\{x_n, \; n \in \mathbb{N}\}$ and $\{y_n, \; n \in \mathbb{N}\}$ be two sequences satisfying large deviations principles on $\mathscr{X}$ and $\mathscr{Y}$, with rate functions I_X and I_Y. Suppose that they are independent, then the sequence $\{(x_n, y_n), \; n \in \mathbb{N}\}$ satisfies a large deviations principle on $\mathscr{X} \times \mathscr{Y}$ with rate function*

$$I_{X,Y}(x, y) = I_X(x) + I_Y(y). \tag{2.11}$$

In this paper, we will establish large deviations principles relying on indirect methods. Once we have a large deviations principle for one sequence of random variables, we

can effortlessly obtain large deviations principles for a whole class of random sequences, namely, those obtained via continuous transformations. We present the *contraction principle*, the tool that enables this.

THEOREM 2.6. *If $\{x_n, n \in \mathbb{N}\}$ satisfies a large deviations principle on $\mathcal{X}$ with rate function I_X, and if $f : \mathcal{X} \to \mathcal{Y}$ is a continuous function, then the sequence $\{y_n, n \in \mathbb{N}\}$ defined by $y_n = f(x_n)$ satisfies a large deviations principle on $\mathcal{Y}$ with rate function*

$$I_Y(y) = \inf_{x, f(x)=y} I_X(x). \tag{2.12}$$

Sometimes f is "almost continuous," that is, the sequence $\{y_n, n \in \mathbb{N}\}$ is close to a sequence $\{f(x_n), n \in \mathbb{N}\}$, where f is continuous.

DEFINITION-PROPOSITION 2.7. *Let $\{x_n, n \in \mathbb{N}\}$ and $\{y_n, n \in \mathbb{N}\}$ be two sequences on $\mathcal{X}$. They are said exponentially equivalent, if for all $\epsilon > 0$,*

$$\limsup_{n \to \infty} \log \mathbb{P}(\|x_n - y_n\| > \epsilon) = -\infty. \tag{2.13}$$

In this case, if $\{x_n, n \in \mathbb{N}\}$ satisfies a large deviation principle, then $\{y_n, n \in \mathbb{N}\}$ satisfies a large deviations principle, and $I_X = I_Y$.

THEOREM 2.8. *If $\{x_n, n \in \mathbb{N}\}$ satisfies a large deviations principle on $\mathcal{X}$ with rate function I_X, and if $\{y_n, n \in \mathbb{N}\}$ and $\{f(x_n), n \in \mathbb{N}\}$ are exponentially equivalent, with $f : \mathcal{X} \to \mathcal{Y}$ continuous, then $\{y_n, n \in \mathbb{N}\}$ satisfies a large deviations principle on $\mathcal{Y}$ with rate function*

$$I_Y(y) = \inf_{x \in \mathcal{X}, f(x)=y} I_X(x). \tag{2.14}$$

For a thorough presentation of large deviations' results and applications, we refer to [5].

3. The model

Let $\mathcal{A} = \{A_n, n \in \mathbb{Z}\}$ be a point process and assume that $A_0 \leq 0 < A_1$ and $A_n < A_{n+1}$, $\forall n \in \mathbb{Z}$. We define the $\mathbb{R}_+^*$-valued sequence of r.v.'s $a = \{a_n, n \in \mathbb{Z}\}$ by $a_n = A_{n+1} - A_n$. Let $s = \{s_n, n \in \mathbb{Z}\}$ be another $\mathbb{R}_+^*$-valued sequence of r.v.'s. The sequences a and s are the *input variables* of the model.

Define the sequence of r.v.'s $\mathcal{D} = \{D_n, n \in \mathbb{Z}\}$ by

$$D_n = \sup_{k \leq n} \left[A_k + \sum_{i=k}^{n} s_i \right]. \tag{3.1}$$

A priori the D_n's are valued in $\mathbb{R} \cup \{+\infty\}$. Assume that a and s are such that the D_n's are almost surely finite. Set $d_n = D_{n+1} - D_n$ and $d = \{d_n, n \in \mathbb{Z}\}$. Define an additional sequence of r.v.'s $r = \{r_n, n \in \mathbb{Z}\}$, valued in $\mathbb{R}_+^*$, by

$$r_n = \min(D_n, A_{n+1}) - A_n. \tag{3.2}$$

The sequences d and r are the *output variables* of the model. In view of the future analysis, it is convenient to introduce the sequence $w = \{w_n, n \in \mathbb{Z}\}$ of random variables valued in $\mathbb{R}_+$, defined by

$$w_n = D_n - s_n - A_n = \sup_{k \leq n-1} \left[\sum_{i=k}^{n-1} (s_i - a_i) \right]^+. \tag{3.3}$$

These random variables satisfy the following recursion (Lindley's equation):

$$w_{n+1} = \left[w_n + s_n - a_n \right]^+. \tag{3.4}$$

Using the variables w_n, we can give alternative definitions of D_n and r_n:

$$\forall l \leq n, \quad D_n = \left[w_l + A_l + \sum_{i=l}^{n} s_i \right] \vee \max_{l < k \leq n} \left[A_k + \sum_{i=k}^{n} s_i \right], \tag{3.5}$$

$$r_n = \min \{w_n + s_n, a_n\} = s_n + w_n - w_{n+1}. \tag{3.6}$$

We now interpret the variables defined above in two different contexts: a *queueing model* and a *storage model*.

3.1. The single server queue.

We are concerned with a single server queue where each customer is characterized by an instant of arrival in the queue and a service demand. Customers are served upon their arrival in the queue and in their order of arrival. Since there is a single server, a customer may have to wait in a buffer before the beginning of its service. Using Kendall's nomenclature, our model is a $\cdot / \cdot /1/\infty/$FIFO queue. The customers are numbered by $\mathbb{Z}$ according to their order of arrival in the queue (customer 1 being the first one to arrive strictly after instant 0). Let A_n be the instant of arrival of customer n and s_n its service time. Then the variables defined in (3.1)–(3.4) have the following interpretations:

    (i) D_n is the instant of departure of customer n from the queue, after completion of its service; $\{d_n, n \in \mathbb{Z}\}$ is the sequence of interdeparture times;

    (ii) w_n is the waiting time of customer n in the buffer between its arrival and the beginning of its service;

    (iii) r_n is the time spent by customer n at the *very back* of the queue.

The variables $\{r_n, n \in \mathbb{Z}\}$ are less classical in queueing theory [16].

3.2. The storage model.

Some product P is supplied, sold, and stocked in a store in the following way: events occur at integer-valued epochs, called *slots*. At each slot, an amount of P is supplied and an amount of P is asked for by potential buyers. The rule is to meet all demands, if possible. The demand of a given slot which is not met *is lost*. The supply of a given slot which is not sold *is not lost* and is stocked for future consideration.

    Let s_n be the amount of P supplied at slot $n+1$, and let a_n be the amount of P asked for at the same slot. In this context, the variables in (3.1)–(3.4) can be interpreted as follows:

    (i) w_n is the level of the stock at the end of slot n. It evolves according to (3.4);

(ii) r_n is the demand met at slot $n + 1$, see (3.6); it is the amount of P departing at slot $n + 1$.

The variables $\{D_n,\ n \in \mathbb{Z}\}$ do not have a natural interpretation in this model.

It is important to remark that while the equations driving the *single server queue* and the *storage model* are exactly the same, the relevant variables are different. The important variables are the ones corresponding to the departures from the system. The departures are coded in the variables $\{d_n,\ n \in \mathbb{Z}\}$ for the single server queue and in the variables $\{r_n,\ n \in \mathbb{Z}\}$ for the storage model. For a more detailed discussion of these two models we refer to [7].

3.3. Rare events. We recall that w_n is the waiting time of customer n before it starts its service (resp., the level of the stock at the end of slot n) and it is given by

$$w_n = \left(w_{n-1} + s_{n-1} - a_{n-1}\right)^+ = \sup_{m \le n}\left[\sum_{k=m}^{n-1} s_k - a_k\right]. \tag{3.7}$$

We assume that the sequences $a = \{a_n,\ n \in \mathbb{Z}\}$ and $s = \{s_n,\ n \in \mathbb{Z}\}$ are stationary and ergodic. Under the stability condition $\mathbb{E}s_0 < \mathbb{E}a_0$, we can use the above expression of w_n to give its asymptotic behavior using the large deviations properties of the input variables $a = \{a_n,\ n \in \mathbb{Z}\}$ and $s = \{s_n,\ n \in \mathbb{Z}\}$. We start with an example.

Example 3.1. Suppose that customers arrive in a deterministic fashion, that is, $a_n = 1$, $\forall n$, and that the sequence $\{s_n,\ n \in \mathbb{Z}\}$ is i.i.d. with

$$\mathbb{P}(s_n = 2) = 1 - \mathbb{P}(s_n = 0) = p < \frac{1}{2}. \tag{3.8}$$

The process $\{w_n,\ n \in \mathbb{Z}\}$ is a discrete time birth-and-death process with stationary distribution

$$\mathbb{P}(w_0 \ge q) = \left(\frac{p}{1 - p}\right)^q. \tag{3.9}$$

We get

$$\frac{1}{n}\log\mathbb{P}(w_0 \ge nq) = -\delta q, \tag{3.10}$$

where $\delta = \log((1 - p)/p)$. This approximation remains valid under general conditions on the input variables, with an expression of δ that depends upon the rate functions associated to the sequences a and s. Indeed it has been proved [2, 8, 12] that the stationary version of w_n satisfies the following property:

$$\lim_{n \to \infty} \frac{1}{n}\log\mathbb{P}(w_0 > nq) = -\delta q, \tag{3.11}$$

or alternatively

$$\mathbb{P}\left(\frac{w_0}{n} > q\right) \asymp e^{-n\delta q}. \tag{3.12}$$

First we give a heuristic proof of this result which will be asserted rigorously in Section 4. We assume the sequences $\{a_n, n \in \mathbb{Z}\}$ and $\{s_n, n \in \mathbb{Z}\}$ i.i.d. mutually independent with

$$\Lambda_A(\theta) = \log \mathbb{E}e^{\theta a_0}, \qquad \Lambda_S = \log \mathbb{E}e^{\theta s_0} \tag{3.13}$$

finite near the origin. Let $x_n = s_n - a_n$ for all $n \in \mathbb{Z}$, and $X_k = \sum_{i=1}^{k} x_{-i}$ for all $k \in \mathbb{N}$ then $w_0 = \sup_{k \geq 0} X_k$, with $X_0 = 0$. Using Cramér's theorem (Theorem 2.4), the sequences $\{a_n, n \in \mathbb{Z}\}$ and $\{s_n, n \in \mathbb{Z}\}$ satisfy large deviations principles on $\mathbb{R}$ with rate function I_A and I_S the Legendre transforms of Λ_A and Λ_S. Thus the sequence $\{x_n, n \in \mathbb{Z}\}$ satisfies a large deviations principle with rate function I_X, which is the Legendre transform of $\Lambda_X(\theta) = \Lambda_S(\theta) + \Lambda_A(-\theta)$, that is, $\mathbb{P}(X_n/n > x) \asymp e^{-nI_X(x)}$. Moreover,

$$\mathbb{P}(w_0 \geq nq) = \mathbb{P}\left(\sup_{k \geq 0} X_k \geq nq\right) = \mathbb{P}\left(\cup_{k \geq 0} \{X_k \geq qn\}\right) \leq \sum_{k=1}^{\infty} \mathbb{P}(X_k \geq qn). \tag{3.14}$$

Since $\mathbb{P}(X_k \geq nq) = \mathbb{P}(X_k/k \geq nq/k) \asymp e^{-kI_X(nq/k)}$, we check that

$$\mathbb{P}(w_0 \geq nq) \asymp \sum_{k=1}^{\infty} e^{-nq(I_X(nq/k)/(nq/k))}. \tag{3.15}$$

We conclude using the principle of the largest term, that is,

$$\sum_{k=1}^{\infty} e^{-nq(I_X(nq/k)/(nq/k))} \asymp e^{-qn\delta}, \tag{3.16}$$

where $\delta = \inf_{x>0}(I_X(x)/x)$ and we get (3.11). The principle of the largest term translates the fact that rare events occur in the most likely way. Indeed, the dominant term in (3.15), which gives the explosion of the waiting times in the queueing system (or the overflow of the stock in the storage model), happens following the most probable scenario.

4. Large deviations for the workload process

First, we give the assumptions under which the approximation (3.11) is fulfilled. For more general assumptions we refer to the paper by Ganesh et al. [11].

Assumptions 4.1. (i) The sequences $\{a_n, n \in \mathbb{Z}\}$ and $\{s_n, n \in \mathbb{Z}\}$ are i.i.d., mutually independent. Their cumulant generating functions are given by

$$\Lambda_A(\theta) = \log \mathbb{E}e^{\theta a_0}, \qquad \Lambda_S(\theta) = \log \mathbb{E}e^{\theta s_0}. \tag{4.1}$$

We assume that both Λ_A and Λ_S are differentiable near the origin.

   (ii) The stability condition

$$\frac{1}{\mu} = \Lambda'_S(0) = \mathbb{E}s_0 < \mathbb{E}a_0 = \Lambda'_A(0) = \frac{1}{\lambda} \tag{4.2}$$

is satisfied.

Under (i) and (ii), we have the following proposition.

PROPOSITION 4.2. *The sequence $\{w_0/n, \ n \in \mathbb{N}^*\}$ satisfies a large deviations principle with good rate function $I_W(q) = \delta q$, that is, $\lim_{n \to \infty}(1/n)\log \mathbb{P}(w_0/n \geq q) = -\delta q$, with*

$$\delta = \inf_{0 < a < s} \frac{I_{A,S}(a,s)}{s-a} = \sup\{\theta : \Lambda_S(\theta) + \Lambda_A(-\theta) \leq 0\}. \tag{4.3}$$

Proof. First we prove (4.3). Recall that $x_n = s_n - a_n$, $X_n = \sum_{i=1}^{n} x_{-i}$, and

$$w_0 = \sup_{k \geq 0} X_k. \tag{4.4}$$

By the independence assumption, we have

$$\Lambda_X(\theta) = \Lambda_S(\theta) + \Lambda_A(-\theta). \tag{4.5}$$

By direct application of the contraction principle, we check that

$$I_X(x) = \inf_{y > x} \{I_S(y) + I_A(y-x)\}. \tag{4.6}$$

Let $\theta \leq \inf_{x \geq 0} I_X(x)/x$, then

$$\begin{aligned}
\theta \leq \inf_{x \geq 0} \frac{I_X(x)}{x} &\Longleftrightarrow \theta \leq \frac{I_X(x)}{x}, \quad \forall x \geq 0 \\
&\Longleftrightarrow \theta x - I_X(x) \leq 0, \quad \forall x \geq 0 \\
&\Longleftrightarrow \sup_{x \geq 0}\{\theta x - I_X(x)\} \leq 0 \\
&\Longleftrightarrow \Lambda_X(\theta) \leq 0.
\end{aligned} \tag{4.7}$$

The last equivalence is due to the fact that $\mathbb{E}x_0 < 0$ (stability condition) and (2.8). We proved that

$$\inf_{0 < a < s} \frac{I_{A,S}(a,s)}{s-a} = \inf_{x \geq 0} \frac{I_X(x)}{x} = \sup\{\theta : \Lambda_S(\theta) + \Lambda_A(-\theta) \leq 0\}. \tag{4.8}$$

(i) *Lower bound:* for $q > 0$, we have $\mathbb{P}(w_0 \geq q) \geq P(X_k \geq q)$. Notice that for $p \geq q/\lceil q/p \rceil$, we have

$$\mathbb{P}(w_0 \geq q) \geq \mathbb{P}\left(\frac{1}{\lceil q/p \rceil} X_{\lceil q/p \rceil} \geq p\right). \tag{4.9}$$

Since $1/q \geq (1/p)(1/\lceil q/p \rceil)$, we get

$$\begin{aligned}
\liminf_{q \to \infty} \frac{1}{q}\log \mathbb{P}(w_0 \geq q) &\geq \frac{1}{p}\liminf_{q \to \infty}\frac{1}{\lceil q/p \rceil}\log \mathbb{P}\left(\frac{1}{\lceil q/p \rceil}X_{\lceil q/p \rceil} \geq p\right) \\
&= \frac{1}{p}\liminf_{n \to \infty}\frac{1}{n}\log \mathbb{P}\left(\frac{1}{n}X_n \geq p\right) \geq -\frac{1}{p}I_X(p),
\end{aligned} \tag{4.10}$$

for all $p > 0$. We conclude that

$$\liminf_{q \to \infty} \frac{1}{q} \log \mathbb{P}(w_0 \geq q)$$

$$\geq -\inf_{x>0} \frac{1}{x} I_X(x) = -\inf_{x>0} \frac{1}{x} \inf_{y>x} \{I_S(y) + I_A(y-x)\} \tag{4.11}$$

$$= -\inf_{0<a<s} \frac{I_S(s) + I_A(a)}{s-a}.$$

(ii) *Upper bound:* recall that $\Lambda_X(\theta) = \Lambda_S(\theta) + \Lambda_A(-\theta)$. By the stability condition, we have

$$\Lambda_X'(0) = \log \frac{\mathbb{E}(s_1)}{\mathbb{E}(a_1)} < 0. \tag{4.12}$$

Moreover, Λ_X is differentiable near 0 with $\Lambda_X(0) = 0$. Thus, there is a constant $\Theta > 0$ such that $\Lambda_X(\Theta) < 0$ and $\mathbb{E}e^{\Theta X_n} < \infty$. Applying Chernoff's bound, we get, for $n \in \mathbb{N}$,

$$\mathbb{P}(X_n \geq q) \leq e^{-\Theta q} \mathbb{E}e^{\Theta X_n}. \tag{4.13}$$

This leads to

$$\limsup_{q \to \infty} \frac{1}{q} \log \mathbb{P}(X_n \geq q) \leq -\Theta. \tag{4.14}$$

For $N \in \mathbb{N}$, we check that

$$\mathbb{P}\left(\max_{0 \leq n \leq N} X_n \geq q\right) \leq N \max_{0 \leq n \leq N} \mathbb{P}(X_n \geq q). \tag{4.15}$$

Allowing q go to infinity,

$$\limsup_{q \to \infty} \frac{1}{q} \log \mathbb{P}\left(\max_{0 \leq n \leq N} X_n \geq q\right) \leq \max_{0 \leq n \leq N} \limsup_{q \to \infty} \log \frac{1}{q} \mathbb{P}(X_n \geq q) \leq -\Theta. \tag{4.16}$$

We need now to have a bound for large values of n. Applying the union bound, we have

$$\mathbb{P}\left(\sup_{n>N} X_n \geq q\right) \leq \sum_{n>N} \mathbb{P}(X_n \geq q) \leq e^{-\Theta q} \sum_{n>N} \mathbb{E}e^{\Theta X_n}. \tag{4.17}$$

Since $(1/n) \log \mathbb{E}e^{\Theta X_n} = \Lambda_X(\Theta) < 0$, there is $0 < \epsilon < -\Lambda_X(\Theta)$ et $N_\Theta \in \mathbb{N}$ such that

$$\forall n > N_\Theta, \quad \frac{1}{n} \log \mathbb{E}e^{\Theta X_n} \leq -\epsilon. \tag{4.18}$$

Combining (4.17) and (2.10), we obtain

$$\mathbb{P}\left(\sup_{n>N_\Theta} X_n > q\right) \leq e^{-\Theta q} \sum_{n>N_\Theta} e^{-n\epsilon} < \frac{e^{-\Theta q}}{1-e^{-\epsilon}} e^{-(N_\Theta+1)\epsilon}. \tag{4.19}$$

To sum up, we derived the following bound:

$$\limsup_{q\to\infty} \frac{1}{q} \log \mathbb{P}\left(\sup_{n>N_\Theta} X_n \ge \right) \le -\Theta. \tag{4.20}$$

Therefore, for $\Theta > 0$ such that $\Lambda_X(\Theta) < 0$, we have

$$\limsup_{q\to\infty} \frac{1}{q} \log \mathbb{P}\left(\sup_{n\ge0} X_n \ge q \right) \le -\Theta. \tag{4.21}$$

We conclude using (4.3). $\qquad\square$

This result can be interpreted naively in terms of the following approximation:

$$\mathbb{P}\left(\sup_n X_n \ge q \right) \approx \sup_n \mathbb{P}(X_n \ge q). \tag{4.22}$$

This is nothing more than a consequence of the fact that rare events occur in the most likely way. Moreover, using this approximation one can predict the frequency at which the waiting time (or the level of stock) overflows a given threshold. Indeed if (3.11) is fulfilled, then we represent the graph (in logarithmic squale) of the frequency of exceeding a level q. The above approximation (**Proposition 4.2**) inspired numerous applications (described by Courcoubetis et al. [3]) of large deviations to the analysis of the statistics of networks of queues.

5. Effective bandwidth

The *bandwidth* of a traffic flow has become a classical feature in communication networks literature. We are interested in the notion of *effective bandwidth* introduced by Kelly [13], which gives an analytical way of describing the properties of a stochastic flow by means of rare events occurring in the network through which it passes.

In practice, the queues and stores we are interested in have finite capacities, that is, the queue (resp., the store) rejects customers (resp., stock) each time the waiting time (resp., the amount of stock) overpasses a given threshold. In the queueing context, we assume $\{s_n, n \in \mathbb{Z}\}$ given and we seek the (deterministic) minimal value a_p of interarrival times such that the probability that the waiting times exceed the threshold is less than a fixed value p. For the storage model, we assume $\{a_n, n \in \mathbb{Z}\}$ given and we seek the maximal value s_p of product arriving at the store at each slot of time, such that the store rejects it with a probability less than p. More precisely, we want to identify the arrivals to both models such that

$$\mathbb{P}(w_0 \ge q) \le p, \tag{5.1}$$

for given values of q and p.

Queue. Let $a_n = a$, $\forall n \in \mathbb{Z}$, that is, $\Lambda_A(\theta) = a\theta$. If s is i.i.d., then by **Proposition 4.2**,

$$\mathbb{P}(w_0 \ge q) \le e^{-\delta(a)q} \tag{5.2}$$

with

$$\delta(a) = \sup\{\theta : \theta \geq 0, \, \Lambda_S(\theta) \leq a\theta\}. \tag{5.3}$$

The minimal interarrival time for the inequality (5.1) to be fulfilled is given by a_p such that

$$a_p = \inf\{a : a \geq 0 \text{ et } e^{-\delta(a)q} \leq p\}. \tag{5.4}$$

Let $\theta_p = -\log p/q$, then $\delta(a_p) = \theta_p$ and

$$a_p = \frac{\Lambda_S(\theta_p)}{\theta_p}. \tag{5.5}$$

The variable a_p is the *effective bandwidth* of the queue. For more details we refer to Kelly's review [14].

Store. Let $s_n = s$, $\forall n \in \mathbb{Z}$, then $\Lambda_S(\theta) = \theta s$ and

$$\delta(s) = -\inf\{\theta < 0 : \Lambda_A(\theta) \leq s\theta\}. \tag{5.6}$$

The inequality (5.1) is satisfied if no more than s_p amount of stock arrives to the store at each slot with

$$s_p = \sup\{s \geq 0 : e^{-\delta(s)q} \leq p\}. \tag{5.7}$$

The quantity s_p is the *effective bandwidth* of the store, with

$$s_p = \frac{\Lambda_A(\theta_p)}{\theta_p}. \tag{5.8}$$

We generalize this notion to nondeterministic arrivals by introducing the functions $\alpha_A(\theta)$ and $\alpha_S(\theta)$ given by

$$\alpha_A(\theta) = \frac{\Lambda_S(\theta)}{\theta}, \qquad \alpha_S(\theta) = \frac{\Lambda_A(\theta)}{\theta}, \tag{5.9}$$

representing the effective bandwidths of the queue and the store, respectively.

6. Principle of large deviations for the output variables

We proved, in Section 4, a large deviations principle for the workload process using classical techniques. It is generally hard to apply these techniques to derive large deviations principles for the output variables. In this section, we will apply the contraction principle to this end. Following the outline of the proof of the existence of fixed points for discrete-time queues in [10], we use theoretical results on large deviations for continuous-time processes.

First, we go back to the workload process to illustrate this method. For $n \in \mathbb{N}^*$, we define

$$A_n = \sum_{i=1}^{n} a_{-i}, \qquad S_n = \sum_{i=1}^{n} s_{-i}. \tag{6.1}$$

It is obvious that $w_0 = \sup_{k \geq 0}(S_k - A_k)$.

To apply the contraction principle it is crucial to define an adapted topology under which the workload process is obtained through a continuous mapping of the input variables. Therefore, we define the polygonal approximation in n of a given sequence $\{X_k, k \in \mathbb{N}^*\}$ by

$$\tilde{X}_n(t) = \frac{1}{n} X_{\lfloor nt \rfloor} + \left(t - \frac{\lfloor nt \rfloor}{n} \right) \left(X_{\lfloor nt \rfloor + 1} - X_{\lfloor nt \rfloor} \right), \quad t \geq 0. \tag{6.2}$$

Let $\tilde{A}_n$ and $\tilde{S}_n$ be the polygonal approximations in n of the sequences $\{A_k, k \in \mathbb{N}\}$ and $\{S_k, k \in \mathbb{N}\}$. Notice that for $n \in \mathbb{N}^*$,

$$\begin{aligned}
\frac{w_0}{n} &= \frac{1}{n} \sup_{k \in \mathbb{N}} (S_k - A_k) = \frac{1}{n} \sup_{t \geq 0} \left(S_{\lfloor nt \rfloor} - A_{\lfloor nt \rfloor} \right) \\
&= \sup_{t > 0} \left(\tilde{S}_n(t) - \tilde{A}_n(t) \right).
\end{aligned} \tag{6.3}$$

A function x is absolutely continuous (on $\mathbb{R}$) if for every $\epsilon > 0$, there is $\eta > 0$ such that for any finitely many disjoint open intervals $(a_i, b_i), i = 1, \dots, n$, satisfying $\sum_{i=1}^{n}(b_i - a_i) \leq \eta$, we have $\sum_{i=1}^{n}(f(b_i) - f(a_i)) \leq \epsilon$. If x is absolutely continuous, then its derivative x' exists almost everywhere and we can write

$$x(v) - x(u) = \int_{u}^{v} x'(t)dt. \tag{6.4}$$

For $\mu > 0$, we define $\mathcal{E}_\mu$ (resp., $\mathcal{A}_\mu$) the set of continuous functions (resp., absolutely continuous) $x : \mathbb{R}_+ \to \mathbb{R}$ with $x(0) = 0$ and

$$\lim_{t \to +\infty} \frac{x(t)}{t+1} = \frac{1}{\mu} < \infty, \tag{6.5}$$

equipped with the norm

$$\|x\| = \sup_{t \in \mathbb{R}_+} \left| \frac{x(t)}{t+1} \right|. \tag{6.6}$$

We now focus on large deviations on processes in continuous time in $\mathcal{E}_\mu$.

Definition 6.1. A sequence of processes $\{X_n, n \in \mathbb{N}\}$, where $X_n \in \mathcal{E}_\mu$, satisfies a *functional large deviations principle with linear geodesics*, with instantaneous rate function I, if
    (i) the function I is a rate function, with $I(1/\mu) = 0$,

(ii) the sequence $\{X_n, n \in \mathbb{N}\}$ satisfies a large deviations principle $\mathscr{E}_\mu$ with rate function

$$\mathscr{I}_X(\phi) = \begin{cases} \int_0^{+\infty} I(\phi'(t))dt & \text{if } \phi \in \mathscr{A}_\mu, \\ +\infty & \text{otherwise.} \end{cases} \tag{6.7}$$

Dembo and Zajic explored extensions [4] of large deviations principles to processes in continuous time under the topology of uniform convergence on compact intervals. However, this topology is not appropriate in the context of queueing systems. We introduce a coarser topology corresponding to the norm (6.6). The following large deviations principle is due to Ganesh and O'Connell [9]. Notice that if $x = \{x_n, n \in \mathbb{N}\}$ an i.i.d. sequence with mean $1/\mu$, then for $n \in \mathbb{N}^*$, $\widetilde{X}_n \in \mathscr{E}_\mu$, where $\widetilde{X}_n$ is the polygonal approximation in n of the sequence $X = \{X_n, n \in \mathbb{N}\}$ with $X_n = \sum_{i=1}^n x_i$

THEOREM 6.2 [9]. *Let $x = \{x_n, n \in \mathbb{N}\}$ be a sequence of real-valued random variables and Λ_X its cumulant generating function, which is differentiable near the origin. Then the sequence of polygonal approximations $\{\widetilde{X}_n, n \in \mathbb{N}^*\}$ of $X = \{X_n, n \in \mathbb{N}\}$, where $X_n = \sum_{i=1}^n x_i$, satisfies a functional large deviations principle with linear geodesics on $\mathscr{E}_\mu$ equipped with the norm (6.6), with mean $1/\mu = \Lambda'_X(0)$ and instantaneous rate function I_X, that is,*

$$\mathscr{I}_X(\phi) = \begin{cases} \int_0^{+\infty} I_X(\phi'(t))dt & \text{if } \phi \in \mathscr{A}_\mu, \\ +\infty & \text{otherwise.} \end{cases} \tag{6.8}$$

Under (i), the processes $\widetilde{A}_n$ and $\widetilde{S}_n$ satisfy functional large deviations principles with linear geodesics and instantaneous rate functions $\mathscr{I}_A$ and $\mathscr{I}_S$ defined by

$$\mathscr{I}_A(\phi) = \begin{cases} \int_0^{+\infty} I_A(\phi'(t))dt & \text{if } \phi \in \mathscr{A}_\lambda, \\ +\infty & \text{otherwise,} \end{cases} \tag{6.9}$$

with $I_A(x) = \sup_{\theta \in \mathbb{R}}\{\theta x - \Lambda_A(\theta)\}$, the Legendre transform of Λ_A. Define I_S and $\mathscr{I}_S$ in the same way.

We check (see [11]) that $f : \mathscr{E}_\mu \times \mathscr{E}_\lambda \to \mathbb{R}_+$ defined by

$$f(\phi, \psi) = \sup_{t>0}\left(\psi(t) - \phi(t)\right) \tag{6.10}$$

is continuous on $\mathscr{E}_\mu \times \mathscr{E}_\lambda$ for the topology induced by the norm (6.6). Indeed f is not continuous under the topology of uniform convergence [11, Example 5.2].

Applying the contraction principle, we check that $\{w_1/n, n \in \mathbb{Z}\}$ satisfies a large deviations principle on $\mathbb{R}_+$ with rate function

$$\begin{aligned} I_W(q) &= \inf\left\{\int_0^\infty I_{A,S}(\phi'(s))ds \mid (\phi_1, \phi_2) \in \mathscr{A}_\lambda \times \mathscr{A}_\mu, f(\phi_1, \phi_2) = q\right\} \\ &= \inf\left\{\int_0^\infty I_{A,S}(\phi'(s))ds \mid (\phi_1, \phi_2) \in \mathscr{A}_\lambda \times \mathscr{A}_\mu, \sup_{t>0}\left(\phi_2(t) - \phi_1(t)\right) = q\right\}, \end{aligned} \tag{6.11}$$

where $I_{A,S}(\phi'(s)) = I_A(\phi_1'(s)) + I_S(\phi_2'(s))$. Combining this with Proposition 4.2, we have the following corollary.

COROLLARY 6.3. *The solution of the following optimization problem:*

$$Minimizing \int_0^\infty I_{A,S}(\phi'(s))\,ds \, ver\, \phi = (\phi_1, \phi_2) \in \mathscr{A}_\lambda \times \mathscr{A}_\mu$$
$$subject\ to\ \sup_{t>0}(\phi_2(t) - \phi_1(t)) = q \tag{6.12}$$

is given by $I_W(q) = \delta q$ where δ is defined in (4.3).

In the following section, we make use of the same techniques to give large deviations principles for the output variables

$$d_n = a_n + w_{n+1} - w_n + s_{n+1} - s_n, \qquad r_n = s_n - w_{n+1} + w_n. \tag{6.13}$$

We define $D = \{D_n,\ n \in \mathbb{N}\}$ and $R = \{R_n,\ n \in \mathbb{N}\}$ by

$$D_n = \sum_{i=1}^n d_{-i}, \qquad R_n = \sum_{i=1}^n r_{-i}. \tag{6.14}$$

Let $\tilde{D}_n$ and $\tilde{R}_n$ be their polygonal approximations in n. Under (i), (ii) and by Loynes' scheme [15], we check that

$$(\tilde{D}_n, \tilde{R}_n) \in \mathscr{E}_\lambda \times \mathscr{E}_\mu. \tag{6.15}$$

6.1. Optimization problem. To make the proof as clear as possible, we introduce a new variable

$$b_n = a_n + w_{n+1} - w_n, \tag{6.16}$$

which stands, in a queueing context, for the time between instances of beginning of services for customers n and $n + 1$. Let $B = \{B_n,\ n \in \mathbb{N}\}$ defined by $B_n = \sum_{i=1}^n b_{-i}$ and $\tilde{B}_n$ its polygonal approximation in n then $\tilde{B}_n \in \mathscr{E}_\lambda$. Under (i), we have the following lemma.

LEMMA 6.4. *For $0 \le t \le 1$ fixed, the sequences $\{\tilde{D}_n(t),\ n \in \mathbb{N}^*\}$ and $\{\tilde{B}_n(t),\ n \in \mathbb{N}^*\}$ are exponentially equivalent.*

Proof. Let $\gamma > 0$,

$$\mathbb{P}\left(\sup_{0 \le t \le 1} |\tilde{D}_n(t) - \tilde{B}_n(t)| > \gamma\right) \le \mathbb{P}\left(\sup_{0 \le k \le n} |s_{-k} - s_0| > n\gamma\right)$$

$$\le \sum_{k=0}^n \mathbb{P}(|s_{-k} - s_0| > n\gamma) \tag{6.17}$$

$$\le e^{-\gamma\nu n}\sum_{k=0}^n \mathbb{E}e^{\gamma|s_{-k}-s_0|} \le (n+1)e^{-\gamma\nu n}e^{2\Lambda_S(\gamma)}.$$

Since $\lim_{n\to\infty}\log((n+1)e^{-\gamma\nu n}) = -\infty$ and thanks to assumption (i) ($\Lambda_S(\gamma) < \infty$ for γ close to 0), we are done. $\qquad\square$

Thus if one of the sequences $\{\tilde{D}_n(t),\ n\in\mathbb{N}^*\}$ and $\{\tilde{B}_n(t),\ n\in\mathbb{N}^*\}$ satisfies a large deviations principle then the other does too and they both have the same rate function. Since the expression of $\tilde{B}_n$ (in terms of the input variables) is easier than the one of $\tilde{D}_n$, we concentrate on $\tilde{B}_n$ and $\tilde{R}_n$. In fact, we have the following.

Lemma 6.5. *The sequences $\tilde{B}_n(t)$ and $\tilde{R}_n(t)$ satisfy*

$$\tilde{B}_n(t) = \tilde{A}_n(t) + \sup_{s>0}\{\tilde{S}_n(s) - \tilde{A}_n(s)\} - \sup_{s>t}\{(\tilde{S}_n(s) - \tilde{S}_n(t)) - (\tilde{A}_n(s) - \tilde{A}_n(t))\},$$
$$\tilde{R}_n(t) = \tilde{S}_n(t) - \sup_{s>0}\{\tilde{S}_n(s) - \tilde{A}_n(s)\} + \sup_{s>t}\{(\tilde{S}_n(s) - \tilde{S}_n(t)) - (\tilde{A}_n(s) - \tilde{A}_n(t))\} \tag{6.18}$$

Proof. We check the result for $\tilde{B}_n(t)$, the proof goes along the same arguments for $\tilde{R}_n(t)$. Since $\tilde{B}_n(t)$ is a linear interpolation of $B_{\lfloor nt\rfloor}/n$, we will prove the result for $B_{\lfloor nt\rfloor}$. First, we check that

$$B_{\lfloor nt\rfloor} = A_{\lfloor nt\rfloor} + w_0 - w_{-\lfloor nt\rfloor}. \tag{6.19}$$

However, we have $w_0 = \sup_{s>0}\{\tilde{S}_n(s) - \tilde{A}_n(s)\}$ then

$$w_{-\lfloor nt\rfloor} = \sup_{s>t}\{(\tilde{S}_n(s) - \tilde{S}_n(t)) - (\tilde{A}_n(s) - \tilde{A}_n(t))\}. \tag{6.20}$$
$\qquad\square$

Theorem 6.6. *Under assumptions (i) and (ii), the sequence $\{(D_n/n, R_n/n, w_0/n),\ n\in\mathbb{N}^*\}$ satisfies a large deviation principle on $\mathbb{R}_+^3$ with rate function*

$$J(x_1, x_2, w) = \inf\Big\{\delta(w - x_1 + x_2) + I_{A,S}(x_2, x_1);\ \inf_C g(q, \tau, v_1, v_2)\Big\}, \tag{6.21}$$

where

$$g(q, \tau, v_1, v_2) = \tau I_{A,S}\left(\frac{x_2 - v_2}{\tau}, \frac{x_2 - v_2 + w}{\tau}\right) + (1-\tau)I_{A,S}\left(\frac{v_1}{1-\tau}, \frac{v_2 - q}{1-\tau}\right) + \delta q, \tag{6.22}$$

with $\delta = \inf_{0<a<s}((I_A(a) + I_S(s))/(s-a))$ and $C = \{(q, \tau, v_1, v_2) \mid q \geq 0,\ 0 \leq \tau \leq 1,\ x_2 - v_2 + w = x_1 - v_1 + q\}$.

Proof. By Loynes [15] and Lemma 6.5, we check that $(\tilde{B}_n, \tilde{R}_n, w_0/n) = \Phi(\tilde{A}_n, \tilde{S}_n)$ with

$$\Phi: \mathscr{E}_\lambda \times \mathscr{E}_\mu \longrightarrow \mathscr{E}_\lambda \times \mathscr{E}_\mu \times \mathbb{R}_+. \tag{6.23}$$

Let $\Phi = (\Phi_1, \Phi_2, \Phi_3)$ and $\phi = (\phi_1, \phi_2) \in \mathscr{E}_\lambda \times \mathscr{E}_\mu$, we have

$$\Phi_1(\phi)(t) = \phi_1(t) + \sup_{s>0}\{\phi_2(s) - \phi_1(s)\} - \sup_{s>t}\{(\phi_2(s) - \phi_2(t)) - (\phi_1(s) - \phi_1(t))\},$$
$$\Phi_2(\phi)(t) = \phi_2(t) - \sup_{s>0}\{\phi_2(s) - \phi_1(s)\} + \sup_{s>t}\{(\phi_2(s) - \phi_2(t)) - (\phi_1(s) - \phi_1(t))\},$$
$$\Phi_3(\phi)(t) = \sup_{s>0}\{\phi_2(s) - \phi_1(s)\}.$$
$$\tag{6.24}$$

Applying the contraction principle, the sequence $\{(\tilde{B}_n, \tilde{R}_n, w_0/n), n \in \mathbb{N}^*\}$ satisfies a large deviations principle with rate function

$$J(\psi_1, \psi_2, w) = \inf\left\{\int_0^\infty I_{A,S}(\phi_1'(s), \phi_2'(s))\,ds \mid \Phi(\phi_1, \phi_2) = (\psi_1, \psi_2, w)\right\}. \tag{6.25}$$

This variational problem is generally hard to solve, we restrict ourselves to $t = 1$. By the law of large numbers, $\{(\tilde{B}_n(1), \tilde{R}_n(1)), n \in \mathbb{N}^*\} = \{(B_n/n, R_n/n), n \in \mathbb{N}^*\}$ converges to the mean values of the output variables, whereas the large deviations principle gives the fluctuations around these values. The sequence $\{(B_n/n, R_n/n, w_0/n), n \in \mathbb{N}^*\}$ satisfies a large deviations principle with rate function

$$J(x_1, x_2, w) = \inf\left\{\int_0^\infty I_{A,S}(\phi_1'(s), \phi_2'(s))\,ds \mid [\Phi(\phi_1, \phi_2)](1) = (x_1, x_2, w)\right\}. \tag{6.26}$$

We introduce the variables q_0 and q_1 given by

$$\begin{aligned}
q_0 &= \sup_{s>0}\{\phi_2(s) - \phi_1(s)\}, \\
q_1 &= \sup_{s>1}\{(\phi_2(s) - \phi_2(1)) - (\phi_1(s) - \phi_1(1))\}.
\end{aligned} \tag{6.27}$$

We rewrite (6.24) as follows

$$x_1 = \phi_1(1) + q_0 - q_1, \quad x_2 = \phi_2(1) - q_0 + q_1, \quad w = q_0. \tag{6.28}$$

To solve this variational problem we condition on the value of q_1 to be equal to q. By the stability condition $\lambda < \mu$, we check that q is finite. Then $J(x_1, x_2, w)$ is the solution of the following optimization problem

$$\begin{aligned}
&\text{Minimize } \int_0^1 I_{A,S}(\phi'(s))\,ds + \int_1^\infty I_{A,S}(\phi'(s))\,ds \quad \text{over } \phi \in \mathscr{A}_\lambda \times \mathscr{A}_\mu,\ q \geq 0 \\
&\text{subject to } q_1 = q, \quad x_1 = \phi_1(1) + q_0 - q_1, \quad x_2 = \phi_2(1) - q_0 + q_1, \quad w = q_0,
\end{aligned} \tag{6.29}$$

where $\phi = (\phi_1(s), \phi_2(s))$ and $I_{A,S}(\phi'(s)) = I_A(\phi_1'(s)) + I_S(\phi_2'(s))$.

Using the auxiliary variable q allows the decomposition of the initial optimization problem into two minimization problems on $s \in [0, 1]$ and $s > 1$ coupled through q. More precisely, we suppose the variables x_1, x_2, and $w = q_0$ given, if we fix $q = q_1$, then $\phi_1(1)$ and $\phi_2(1)$ are fixed. For $s > 1$, the only constraint on ϕ_1 and ϕ_2 is $q_1 = q$. Consequently, the first minimization consists of

$$\text{Minimizing } \int_1^\infty I_{A,S}(\phi_1'(s), \phi_2'(s))\,ds, \quad \text{subject to } q_1 = q. \tag{6.30}$$

The second minimization consists of

$$\begin{aligned}
&\text{Minimizing } \int_0^1 I_{A,S}(\phi_1'(s), \phi_2'(s))\,ds \\
&\text{subject to } x_1 = \phi_1(1) + q_0 - q_1, \quad x_2 = \phi_2(1) - q_0 + q_1, \quad w = q_0,
\end{aligned} \tag{6.31}$$

knowing $q_1 = q$. In both cases the minimization is for $\phi \in \mathscr{A}_\lambda \times \mathscr{A}_\mu$.

For the first problem, let $\psi \in \mathcal{A}_\lambda \times \mathcal{A}_\mu$ with $\psi_i(s-1) = \phi_i(s) - \phi_i(1)$, $s \geq 1$, $i = 1,2$. By Corollary 6.3, this optimization gives the rate function associated with w_0 given in Proposition 4.2 and the minimum is δq where $\delta = \inf_{a<s}(I_{A,S}(a,s)/(s-a))$.

Hence, to determine $J(x_1, x_2, w)$ it remains to

$$\text{Minimize } \int_0^1 I_A(\phi_1'(s)) + I_S(\phi_2'(s))\,ds + \delta q, \tag{6.32}$$

for $(\phi_1, \phi_2) \in \mathcal{A}_\lambda \times \mathcal{A}_\mu$ subject to

$$q \geq 0, \qquad w = \sup_{s \geq 0}\{\phi_2(s) - \phi_1(s)\},$$
$$x_1 = \phi_1(1) + w - q \tag{6.33}$$
$$x_2 = \phi_2(1) - w + q.$$

(i) Suppose $\sup_{s \geq 0}\{\phi_2(s) - \phi_1(s)\} = \sup_{s \geq 1}\{\phi_2(s) - \phi_1(s)\}$, then $w = q + \phi_2(1) - \phi_2(1)$, that is,

$$x_1 = \phi_2(1), \quad x_2 = \phi_1(1), \quad q = w - x_1 + x_2. \tag{6.34}$$

The optimization problem (6.32) is fulfilled by linear functions ϕ_1 and ϕ_2 and we have

$$\delta(w - x_1 + x_2) + I_A(x_2) + I_S(x_1). \tag{6.35}$$

(ii) If $\phi_2(s) - \phi_1(s)$ reaches its maximum for $\tau \in [0,1]$, then

$$x_1 = -q + \phi_1(1) - \phi_1(\tau) + \phi_2(\tau),$$
$$x_2 = q + \phi_2(1) - \phi_2(\tau) + \phi_1(\tau). \tag{6.36}$$

We define $y_i = \phi_i(\tau)/\tau$, $z_i = (\phi_i(1) - \phi_i(\tau))/(1 - \tau)$, $i = 1,2$ and $\chi \in \mathcal{A}_\lambda \times \mathcal{A}_\mu$,

$$\chi_i(t) = \begin{cases} y_i t & \text{if } t \in]0,\tau], \\ y_i \tau + z_i(t - \tau) & \text{if } t \in]\tau,1]. \end{cases} \tag{6.37}$$

Using the convexity of I_A and I_S and Jensen's inequality, we have

$$\int_0^1 I_{A,S}(\chi_1'(s), \chi_2'(s))\,ds \leq \int_0^1 I_{A,S}(\phi_1'(s), \phi_2'(s))\,ds. \tag{6.38}$$

Since $\int_0^1 I_{A,S}(\chi_1'(s), \chi_2'(s))\,ds = \tau I_{A,S}(y) + (1-\tau)I_{A,S}(z)$ where $y = (y_1, y_2)$ and $z = (z_1, z_2)$, it remains to minimize

$$\delta q + \tau I_{A,S}(y) + (1 - \tau)I_{A,S}(z), \tag{6.39}$$

subject to

$$
\begin{aligned}
q &\geq 0, \quad y_1 \geq y_2, \\
x_1 &= -q + (1-\tau)z_1 + \tau y_2, \\
x_2 &= q + (1-\tau)z_2 \leq (1-\tau)z_1, \\
q &+ (1-\tau)z_2 \leq (1-\tau)z_1.
\end{aligned}
\tag{6.40}
$$

Let $v_1 = (1-\tau)z_1$ and $v_2 = (1-\tau)z_2 + q$, we check that $y_1 = (x_2 - v_2)/\tau$, $y_2 = (w + x_2 - v_2)/\tau$ and

$$
J(x_1, x_2, w) = \inf \left\{ \delta(w - x_1 + x_2) + I_{A,S}(x_2, x_1), \inf_C g(q, \tau, v_1, v_2) \right\},
\tag{6.41}
$$

g and C are defined in Theorem 6.6. $\qquad\square$

Let us give an interpretation of this result in a queueing context. Indeed, the most likely ways to get departures with mean x_1, times spent at the back of queue with mean x_2, and waiting times with mean w are when

(i) all customers belong to the same busy period (the system is never empty). Thus the mean of the departures is equal to the mean of the services and the average time spent at the very back of the queue is equal to the mean of the arrivals,

(ii) or let $(q, \tau, v_1, v_2) \in C$ the variables where g reaches its minimum. Then the customer $-n\tau$ finds the system empty, and the optimal path splits into two periods. Customer $-n$ waits nq time before starting its service. During the first period the average of arrivals and services is $(x_2 - v_2)/\tau$ and $(x_2 - v_2 + w)/\tau$ and it consists of $n\tau$ customers. During the second period the mean of services is $(v_2 - q)/(1 - \tau)$ and the arrivals occur with mean $v_1/(1 - \tau)$.

Equation (6.21) determines the rate functions of the output variables in terms of the rate functions of the input variables. A natural question is when are these functions equal? More precisely, when do we have $(I_D, I_R) = (I_A, I_S)$? We start with an example which translates the result stated in [7] in terms of large deviations.

Example 6.7 (the M/M/1 queue). Suppose that $\{a_n, n \in \mathbb{Z}\}$ and $\{s_n, n \in \mathbb{Z}\}$ are i.i.d. mutually independent and for $t \geq 0$,

$$
P(a_0 > t) = e^{-\lambda t}, \qquad P(s_0 > t) = e^{-\mu t}.
\tag{6.42}
$$

For $\theta < \lambda < \mu$,

$$
\Lambda_A(\theta) = \log\left(\frac{\lambda}{\lambda - \theta}\right), \qquad \Lambda_S(\theta) = \log\left(\frac{\mu}{\mu - \theta}\right).
\tag{6.43}
$$

Since $\delta = \sup\{\theta > 0, \Lambda_S(\theta) + \Lambda_A(-\theta) < 0\}$, we check that $\delta = \mu - \lambda$. Moreover we have

$$
I_A(a) = \lambda a - \log(\lambda a) - 1, \qquad I_S(s) = \mu s - \log(\mu s) - 1.
\tag{6.44}
$$

Resolving the above optimization problem, we have

$$\begin{aligned} J(x_1, x_2, w) &= \lambda x_1 - \log(\lambda x_1) - 1 + \mu x_2 - \log(\mu x_2) - 1 + (\mu - \lambda)w \\ &= I_{A,S}(x_1, x_2) + \delta w. \end{aligned} \tag{6.45}$$

6.2. Existence of fixed points. Suppose $\{a_n, \ n \in \mathbb{Z}\}$ and $\{s_n, \ n \in \mathbb{Z}\}$ i.i.d., mutually independent with means $1/\lambda$ and $1/\mu$. They satisfy large deviations principles with rate functions I_A and I_S differentiable on $\mathbb{R}_+$. Let $\beta > 0$ such that

$$I_S'(x) - I_A'(x) = \beta, \quad \forall x \in \mathbb{R}_+. \tag{6.46}$$

We can rewrite this relation in the following which will be useful for future analysis:

$$I_A(x) - I_A(y) = I_S(x) - I_S(y) - \beta(x - y), \quad \forall (x, y) \in (\mathbb{R}_+)^2. \tag{6.47}$$

Before stating the fixed point result for the Queue/Store operator for the rate functions, we give an interpretation of the relations (6.46) and (6.47) in terms of *exponential tilting* of measures.

Definition 6.8. Let $\beta > 0$, it is said that a measure ν' is the β-exponentially tilted measure of a given measure ν if for all $x \in \mathbb{R}$,

$$\frac{d\nu'}{d\nu}(x) = \frac{e^{\beta x}}{\int e^{\beta x} d\nu(x)}, \tag{6.48}$$

where $(d\nu'/d\nu)(x)$ denotes the density of ν' with respect to ν.

PROPOSITION 6.9. *Suppose $\{a_n, \ n \in \mathbb{Z}\}$ and $\{s_n, \ n \in \mathbb{Z}\}$ satisfy relation (6.46) or (6.47), with marginals ν_A and ν_S. Then ν_A is the β-exponentially tilted function of ν_S.*

Proof. Recall that $I_A(1/\lambda) = 0$. We check that

$$\begin{aligned} \Lambda_A(\theta) &= \sup_{x \in \mathbb{R}} \left[\theta x - I_A(x) \right] = \sup_{x \in \mathbb{R}} \left[\theta x - I_S(x) + I_S\left(\frac{1}{\lambda}\right) + \beta\left(x - \frac{1}{\lambda}\right) \right] \\ &= \sup_{x \in \mathbb{R}} \left[(\theta + \beta)x - I_S(x) \right] - \left[\frac{\beta}{\lambda} - I_S\left(\frac{1}{\lambda}\right) \right] = \Lambda_S(\theta + \beta) - \Lambda_S(\beta). \end{aligned} \tag{6.49}$$

Considering the exponentials on both sides, we have

$$\int_{\mathbb{R}_+} e^{\theta x} d\nu_A(x) = \frac{\int_{\mathbb{R}_+} e^{(\theta + \beta)x} d\nu_S(x)}{\int_{\mathbb{R}_+} e^{\beta x} d\nu_S(x)}. \tag{6.50}$$

$\square$

LEMMA 6.10. *Assume that $\{a_n, \ n \in \mathbb{Z}\}$ and $\{s_n, \ n \in \mathbb{Z}\}$ satisfy assumptions (i) and (ii). Then the tilt coefficient β is given by*

$$\beta = \sup \{\theta : \Lambda_S(\theta) + \Lambda_A(-\theta) \le 0\}. \tag{6.51}$$

Proof. Let $\theta = -\beta$ in (6.49), we have

$$\Lambda_A(-\beta) = \Lambda_S(0) - \Lambda_S(\beta). \tag{6.52}$$

Since $\Lambda_S(0) = 0$, we obtain $\Lambda_A(-\beta) + \Lambda_S(\beta) = 0$. Thanks to (4.3), we have

$$\beta \leq \sup\{\theta \geq 0, \ \Lambda_S(\theta) + \Lambda_A(-\theta) \leq 0\} = \delta. \tag{6.53}$$

The function $G(\theta) = \Lambda_S(\theta) + \Lambda_A(-\theta)$ is convex with $G(0) = 0$. Moreover, by assumption (ii),

$$G'(0) = \Lambda_S'(0) - \Lambda_A'(0) < 0. \tag{6.54}$$

This means that G is equal to zero twice on $[0,\beta]$, at 0 and β, then

$$\beta = \sup\{\theta \geq 0, \ \Lambda_S(\theta) + \Lambda_A(-\theta) \leq 0\} = \delta. \tag{6.55}$$
□

THEOREM 6.11. *Let $\delta > 0$ and the sequences $\{a_n, \ n \in \mathbb{Z}\}$ and $\{s_n, \ n \in \mathbb{Z}\}$ the input variables of the model with ν_A the δ-tilted measure of ν_S (or ν_S the $(-\delta)$-tilted measure of ν_A). Then the sequence $\{(D_n/n, R_n/n, w_0/n), \ n \in \mathbb{N}^*\}$ satisfies a large deviations principle on $\mathbb{R}_+^3$ with rate function*

$$J(x_1, x_2, w) = \delta w + I_{A,S}(x_1, x_2). \tag{6.56}$$

Moreover,

$$I_D = I_A, \qquad I_R = I_S. \tag{6.57}$$

Proof. By Theorem 6.6, the sequence $\{(\tilde{D}_n(1), \tilde{R}_n(1), w_1/n), \ n \in \mathbb{N}^*\}$ satisfies a large deviations principle on $\mathbb{R}_+^3$ with rate function

$$J(x_1, x_2, w) = \inf\left\{\delta(w - x_1 + x_2) + I_{A,S}(x_2, x_1); \ \inf_C g(q, \tau, \nu_1, \nu_2)\right\}. \tag{6.58}$$

We start with the first term of this optimization problem, that is,

$$x_1 = \phi_2(1), \quad x_2 = \phi_2(1), \quad q = w - x_1 + x_2. \tag{6.59}$$

We have $w = x_1 - x_2 + q$, and thanks to (6.47),

$$
\begin{aligned}
\delta(w - x_1 + x_2) + I_{A,S}(x_2, x_1) &= \delta(w - x_1 + x_2) + I_{A,S}(x_1, x_2) + \delta(x_1 - x_2) \\
&= \delta w + I_{A,S}(x_1, x_2).
\end{aligned} \tag{6.60}
$$

For the second term, let $(q, \tau, \nu_1, \nu_2) \in C$,
    (i) if $\nu_1 = \nu_2$, from the convexity property, then

$$
\begin{aligned}
g(q, \tau, \nu_1, \nu_2) &= \delta q + \tau I_{A,S}\left(\frac{x_2 - \nu_2}{\tau}, \frac{x_2 - \nu_2 + w}{\tau}\right) + (1-\tau) I_{A,S}\left(\frac{\nu_2}{1-\tau}, \frac{\nu_2 - q}{1-\tau}\right) \\
&\geq \delta q + I_{A,S}(x_2, x_2 + w - q) = \delta w + I_{A,S}(x_1, x_2);
\end{aligned} \tag{6.61}
$$

(ii) if $v_1 > v_2$, then

$$I_{A,S}\left(\frac{x_2 - v_2}{\tau}, \frac{x_2 - v_2 + w}{\tau}\right) = I_{A,S}\left(\frac{x_2 - v_2 + w}{\tau}, \frac{x_2 - v_2}{\tau}\right) + \delta\frac{w}{\tau}. \tag{6.62}$$

Notice that

$$I_S\left(\frac{v_2 - q}{1 - \tau}\right) = I_S\left(\frac{v_2}{1 - \tau}\right) + I_A\left(\frac{v_2 - q}{1 - \tau}\right) - I_A\left(\frac{v_2}{1 - \tau}\right) - \delta\frac{q}{1 - \tau}. \tag{6.63}$$

Since I_A is convex and $v_1 \geq v_2$, we have

$$I_A\left(\frac{v_2 - q}{1 - \tau}\right) - I_A\left(\frac{v_2}{1 - \tau}\right) \geq I_A\left(\frac{v_1 - q}{1 - \tau}\right) - I_A\left(\frac{v_1}{1 - \tau}\right). \tag{6.64}$$

Thus,

$$I_S\left(\frac{v_2 - q}{1 - \tau}\right) + I_A\left(\frac{v_1}{1 - \tau}\right) \geq I_S\left(\frac{v_2}{1 - \tau}\right) + I_A\left(\frac{v_1 - q}{1 - \tau}\right) - \delta\frac{q}{1 - \tau}. \tag{6.65}$$

Since I_A and I_S are convex and $x_2 - v_2 + w + v_1 - q = x_1$, we have

$$g(q) \geq \delta w + I_{A,S}(x_1, x_2). \tag{6.66}$$

We conclude that $J(x_1, x_2, w) = \delta w + I_{A,S}(x_1, x_2)$. $\qquad\square$

In other words, the above theorem states that if the input variables satisfy large deviations principles with rate functions I_A and I_S with v_A the δ-tilted measure of v_S, then the output variables satisfy large deviations principles with the same rate functions.

Acknowledgments

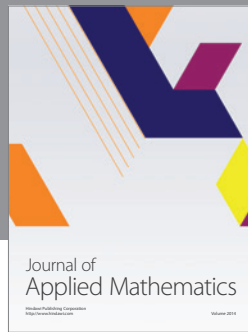## References

[1] P. J. Burke, *The output of a queuing system*, Operations Research **4** (1956), 699–704 (1957).

[2] C.-S. Chang, *Stability, queue length, and delay of deterministic and stochastic queueing networks*, IEEE Transactions on Automatic Control **39** (1994), no. 5, 913–931.

[3] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R. Weber, *Admission control and routing in ATM networks using inference from measured buffer occupancy*, IEEE Transations on Communications **43** (1995), no. 234, 1778–1784.

[4] A. Dembo and T. Zajic, *Large deviations: from empirical mean and measure to partial sums process*, Stochastic Processes and Their Applications **57** (1995), no. 2, 191–224.

[5] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett, Massachusetts, 1993.

[6] M. Draief, J. Mairesse, and N. O'Connell, *Joint Burke's theorem and RSK representation for a queue and a store*, Discrete Random Walks (DRW '03) (C. Banderier and C. Krattenthaler, eds.), vol. AC, Institut Henri Poincaré, Paris, 2003, pp. 69–82.

[7] ———, *Queues, stores, and tableaux*, Journal of Applied Probability **42** (2005), no. 4, 1145–1167.

[8] N. G. Duffield and N. O'Connell, *Large deviations and overflow probabilities for the general single-server queue, with applications*, Mathematical Proceedings of the Cambridge Philosophical Society **118** (1995), no. 2, 363–374.

[9] A. Ganesh and N. O'Connell, *A large deviation principle with queueing applications*, Stochastics and Stochastics Reports **73** (2002), no. 1-2, 25–35.

[10] A. Ganesh, N. O'Connell, and B. Prabhakar, *Invariant rate functions for discrete-time queues*, Annals of Applied Probability **13** (2003), no. 2, 446–474.

[11] A. Ganesh, N. O'Connell, and D. Wischik, *Big Queues*, Lecture Notes in Mathematics, vol. 1838, Springer, Berlin, 2004.

[12] P. W. Glynn and W. Whitt, *Logarithmic asymptotics for steady-state tail probabilities in a single-server queue*, Journal of Applied Probability **31A** (1994), 131–156.

[13] F. Kelly, *Effective bandwidths at multi-class queues*, Queueing Systems. Theory and Applications **9** (1991), no. 1-2, 5–15.

[14] ———, *Notes on effective bandwidth*, Stochastic Networks (F. Kelly and I. Ziedens, eds.), Oxford University Press, Oxford, 1996, pp. 141–168.

[15] R. M. Loynes, *The stability of a queue with non-independent interarrival and service times*, Proceedings of the Cambridge Philosophical Society **58** (1962), 497–520.

[16] B. Prabhakar and R. Gallager, *Entropy and the timing capacity of discrete queues*, IEEE Transactions on Information Theory **49** (2003), no. 2, 357–370.

Moez Draief: Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, UK
E-mail address: m.draief@statslab.cam.ac.uk